Overview of the TREC 2012 Crowdsourcing Track

Mark D. Smucker[1], Gabriella Kazai[2], and Matthew Lease[3]

[1]Department of Management Sciences, University of Waterloo
[2]Microsoft Research, Cambridge, UK
[3]School of Information, University of Texas at Austin

Abstract

In 2012, the Crowdsourcing track had two separate tasks: a text relevance assessing task (TRAT) and an image relevance assessing task (IRAT). This track overview describes the track and provides analysis of the track's results.

1 Introduction

2012 was the second year of the TREC Crowdsourcing track. Track goals for the first two years included: building awareness and expertise with crowdsourcing in the IR community, developing and evaluating new methodologies for crowdsourced search evaluation on a shared task and data set, and creating reusable resources to benefit future IR community experimentation.

The first year was explicitly focused on crowdsourcing. In 2012, we decided to loosen the crowdsourcing requirements and instead focus on a goal of obtaining quality relevance judgments by any means. Any crowdsourcing approach, paid or non-paid, and any platform or home-grown system could be used to obtain the relevance labels, as well as hybrid or fully-automatic methods. The advantage of this change was to give groups freedom in the creation of their solutions. In addition, in year two, we went to simpler data collections and also increased the scale of judgments required by nearly a factor of 10, which brought the task's scale to be much more representative of the type of challenge crowdsourcing methods face. The open-ended task and increased scale has led to innovative attempts at obtaining high quality relevance judgments at low cost, and many of the participating groups have chosen to break new ground with the combination of machine learning and human relevance judgments.

The track consisted of two tasks. One track was a text relevance assessing task (TRAT) and the other was an image relevance assessing task (IRAT). Seven groups participated in TRAT and 2 groups participated in IRAT. We next separately describe each of the tasks, their data, evaluation methods, and results.

2 TRAT

The text relevance assessing task (TRAT) was one of the two TREC 2012 Crowdsourcing Track tasks. The goal of the task was to evaluate approaches to text relevance assessing. We assumed that many participating groups would utilize traditional crowdsourcing platforms, such as Amazon Mechanical Turk, to do the relevance assessing, but the task was open to all approaches that followed the task's guidelines.

The TRAT required participating groups to simulate the relevance assessing role of NIST for 10 of the TREC 8 [11] ad-hoc topics. A key difference between NIST and the participating groups is that the NIST assessors created the search topics and then determined the relevance of documents, while TRAT groups had to rely on their interpretation of the search topic's description and narrative to determine the relevance of a document. While the topic narrative aims to capture a description of what is relevant, it is always possible that the NIST assessors made relevance judgments based on notions of relevance known only to them.

Participating groups had to submit a binary relevance judgment for every document in the judging pools of the ten topics. The submission of a probability of relevance (1.0 means relevant and 0.0 means non-relevant) was optional, but if it was submitted, had to be submitted for all documents in a run.

We placed no limits on the methods that could be employed to obtain relevance judgments. The only restriction was that the NIST judgments (qrels) for TREC 8 and later tracks that used the TREC 8 topics (401-450) were forbidden for use in any way, shape, or manner. For example, participants were forbidden

Report Documentation Page

Form Approved
OMB No. 0704-0188

Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

1. REPORT DATE **NOV 2012**	2. REPORT TYPE	3. DATES COVERED **00-00-2012 to 00-00-2012**

4. TITLE AND SUBTITLE **Overview of the TREC 2012 Crowdsourcing Track**	5a. CONTRACT NUMBER
	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **University of Texas at Austin,School of Information,1616 Guadalupe,Austin,TX,78701**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR’S ACRONYM(S)
	11. SPONSOR/MONITOR’S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

Presented at the Twenty-First Text REtrieval Conference (TREC 2012) held in Gaithersburg, Maryland, November 6-9, 2012. The conference was co-sponsored by the National Institute of Standards and Technology (NIST) the Defense Advanced Research Projects Agency (DARPA) and the Advanced Research and Development Activity (ARDA). U.S. Government or Federal Rights License

14. ABSTRACT

In 2012, the Crowdsourcing track had two separate tasks: a text relevance assessing task (TRAT) and an image relevance assessing task (IRAT). This track overview describes the track and provides analysis of the track’s results.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**9**	

from even using the qrels to help them understand the topic or train workers. Nor could participants make simple counts of the number of relevant documents etc. For training and other purposes, we suggested that participants use the TREC 7 ad-hoc topics and qrels. TREC 7 ad-hoc used the same document collection and had similar topics.

2.1 Data

TRAT used the TREC 8 ad-hoc track as its source of data. We randomly selected 10 topics from the 50 topics used in TREC 8. The 10 topics selected were: 411, 416, 417, 420, 427, 432, 438, 445, 446, and 447. For these topics, we selected all topic-document pairs that NIST had judged in the qrels. In total, 18260 topic-document pairs needed to be judged. These topic-document pairs represented the "test set" for the TRAT.

TRAT used the existing topic descriptions. The topic's title, description, and narrative when taken together defined what the original NIST assessor considered to be relevant and non-relevant to the topic. We clarified to the track participants that the narrative was to take precedence over the description and title, but did not provide further guidance on how to determine relevance other than to provide copies of relevant portions of the TREC 7 and 8 judging guidelines provided to the NIST assessors.

The documents to be judged came from the corpora used by TREC 7 and 8: the Text Research Collection Volumes 4 (May 1996) and 5 (April 1997) minus the Congressional Record (CR). This collection is made available for free to registered TREC participants. The documents come from the following subcollections of volumes 4 and 5:

1. Financial Times, 1991-1994, (FT)
2. Federal Register, 1994 (FR94)
3. Foreign Broadcast Information Service (FBIS)
4. Los Angeles Times (LA).

We advised participants that they could make use of the submitted runs to TREC 8. These same runs were used by NIST to form the judging pools.

2.2 Adjudication

Each participating group designated one of their runs as a primary run for use in a majority vote consensus process. All other runs from a group were secondary runs. The majority vote of the submitted runs was compared to the NIST relevance judgment. When the majority vote differed from the NIST judgment, we adjudicated the final relevance judgment for a document. In total, 459 documents needed adjudication.

Mark Smucker's research group at the University of Waterloo performed the adjudication of the 459 documents. The University of Waterloo did not participate in the track. Mark and two graduate research assistants, Gaurav Baruah and Le Li, separately judged each of the 459 documents. Prior to judging the documents, each judge practiced making judgments for the topic on up to 10 documents (half relevant, half non-relevant) where the majority consensus had strongly agreed with the NIST assessor. For each document, the judges recorded their decision, and for relevant documents marked the location of relevant material, and for non-relevant document, the judges recorded a written reason why the document was not relevant. Of the 459 documents, the judges did not have full agreement on the relevance or differed from the NIST assessor on 270 documents. For these 270 documents, the three judges sat down together and reviewed the documents to reach a final decision. To aid our review, we created a separate system to view our individual judgments on documents and then select a final verdict. For the first 31 judgments we made, the NIST qrel was displayed to us, and fearing that it might bias us, we removed the display of the NIST qrel from the remaining documents.

Table 1 shows some basic statistics about the adjudication. One issue with the adjudication process was that some of the runs were very conservative in their judgments and had very low true positive rates. When we took a majority vote of the runs' judgments, the result is that the majority vote is quite conservative and most disagreements with NIST occur on documents NIST said were relevant but which the majority vote said were non-relevant. Overall, we found NIST to be correct in most cases. Table 2 shows a breakdown of the number of times we reversed a NIST qrel. When a qrel was reversed, it could have been because of a variety of reasons. In some cases, we could find no reason a document was to be considered relevant, and in other cases it might have been that the document did not appear relevant given the search topic. In many ways, the adjudication acts as a means to create a set of qrels that reflects the standard attainable given a third party reading of the search topic. The reality is that only the NIST assessor would be able to know which "mistakes" were actual mistakes and which were times when the assessor's notion of relevance simply failed to be captured by the search topic's description and narrative.

As we did the adjudication, certain topics such as

Topic Title	Topic	#Docs	NIST Rel	%Rel	NumAdj	%Adj	%Adj NIST Rel
salvaging, shipwreck, treasure	411	2056	27	1%	15	1%	80%
Three Gorges Project	416	1235	42	4%	17	1%	65%
measuring creativity	417	2992	75	3%	60	2%	87%
carbon monoxide poisoning	420	1136	33	3%	23	2%	78%
UV damage, eyes	427	1528	50	3%	42	3%	83%
profiling, motorists, police	432	2503	28	1%	34	1%	76%
tourism, increase	438	1798	173	11%	118	7%	84%
women clergy	445	1404	62	5%	29	2%	69%
tourists, violence	446	2020	162	9%	119	6%	90%
Stirling engine	447	1588	16	1%	2	0%	100%

Table 1: Each of the 10 TRAT topics were adjudicated. The columns shown from left to right are the topic title, topic number, total number of documents judged by participants, number of documents judged relevant by NIST, the prevalence of NIST relevant documents, the number of documents adjudicated, the percent of documents adjudicated, and the fraction of adjudicated documents that were NIST relevant.

432 and 446 were quite difficult to judge given the topic's description and narrative. We recommend in the future that some sort of trial adjudication take place prior to the release of topics for crowdsourcing tracks. Such trial adjudication efforts could identify problematic topics that might be considered inappropriate for crowdsourcing given their descriptions and narratives. It might be possible on examination of an assessor's qrels to adjust the topic's narrative. Alternatively, crowdsourcing tracks could consider adopting the notion of a topic authority as found in the TREC Legal track [4].

2.3 Evaluation

All runs are evaluated against the adjudicated qrels. For each topic, the performance of a submitted run is judged on both its binary judgments and its probabilities of relevance. The binary judgments were evaluated using the logistic average misclassification rate (LAM) developed for the Spam Track [2]:

$$LAM = logit^{-1}\left(\frac{logit(fpr) + logit(fnr)}{2}\right), \quad (1)$$

where relevant documents are the positive class, fpr is the false positive rate, fnr is the false negative rate, and

$$logit(p) = \log\frac{p}{1-p}, \quad (2)$$

and

$$logit^{-1} = \frac{e^x}{1+e^x}. \quad (3)$$

Topic	#adj	Rel-to-Non	Non-to-Rel	Total
411	15	1	1	2
416	17	0	3	3
417	60	3	3	6
420	23	1	5	6
427	42	14	1	15
432	34	7	1	8
438	118	16	5	21
445	29	3	1	4
446	119	15	9	24
447	2	0	0	0
Total	459	60	29	89

Table 2: This table shows the number of adjudicated documents per topic and the number of NIST relevance judgments that were reversed as part of the adjudication. For example, topic 416 had 17 documents adjudicated, 0 relevant to non-relevant reversals, 3 non-relevant to relevant reversals, and in total 3 reversals.

We smoothed the calculation of fpr and fnr:

$$fpr = \frac{|FP| + 0.5}{|FP| + |TN| + 1}, \quad (4)$$

$$fnr = \frac{|FN| + 0.5}{|FN| + |TP| + 1}, \quad (5)$$

where $|FP|$ is the number of false positives, etc, as given in Table 3.

We measured the performance of the runs given their probabilities of relevance with the area under the ROC curve (AUC). Only runs that provided probabilities were evaluated using AUC.

	Adjudicated Standard	
Run Submission	Relevant (Pos.)	Non-Relevant (Neg.)
Relevant	TP = True Pos.	FP = False Pos.
Non-Relevant	FN = False Neg.	TN = True Neg.

Table 3: Confusion Matrix. “Pos.” and “Neg.” stand for “Positive” and “Negative” respectively.

Run	TPR	FPR	TNR	FNR	LAM
UIowaS02r	0.772	0.009	0.991	0.228	0.05
SSEC3excl	0.705	0.019	0.981	0.295	0.07
INFLB2012	0.036	0.002	0.998	0.964	0.13
NEUElo3	0.212	0.014	0.986	0.788	0.18
yorku12cs03	0.710	0.174	0.826	0.290	0.22
BUPTPRISZHS	0.211	0.020	0.980	0.789	0.23
OrcVBW16Conf	0.751	0.309	0.691	0.249	0.26

Table 4: Top runs from each participating group ordered by LAM (lower is better) with official smoothing of rates as per Equations 4 and 5. Also shown are the true positive rate (TPR), false positive rate (FPR), true negative rate (TNR), and false negative rate (FNR).

Run	TPR	FPR	TNR	FNR	LAM	AUC
SSEC3inclML	0.777	0.024	0.976	0.223	0.07	0.91
OrcVB1	0.652	0.294	0.706	0.348	0.31	0.81
NEUNugget12	0.299	0.026	0.974	0.701	0.21	0.75
yorku12cs03	0.710	0.174	0.826	0.290	0.22	0.48

Table 5: Top runs from each participating group ordered by AUC (higher is better). Only runs that submitted probabilities of relevance were evaluated using AUC. Also shown are the true positive rate (TPR), false positive rate (FPR), true negative rate (TNR), false negative rate (FNR), and the logistic average misclassification (LAM) rate.

Run	TPR	FPR	FNR	TNR	LAM
UIowaS02r	0.78	0.01	0.22	0.99	0.05
SSEC3excl	0.71	0.02	0.29	0.98	0.07
yorku12cs03	0.72	0.17	0.28	0.83	0.22
NEUNugget12	0.29	0.03	0.71	0.97	0.22
BUPTPRISZHS	0.21	0.02	0.79	0.98	0.25
OrcVBW16Conf	0.76	0.31	0.24	0.69	0.25
INFLB2012	0.02	0.00	0.98	1.00	0.30

Table 6: Top runs from each participating group ordered by LAM (lower is better) using rates smoothed as per Equations 6 and 7. Also shown are the true positive rate (TPR), false positive rate (FPR), true negative rate (TNR), and false negative rate (FNR).

2.4 Results

Seven groups submitted 33 runs, and of these 33 runs, 28 runs also had probabilities of relevance. The seven groups and their run prefixes were:

- Northeastern University [1], runs: NEU*
- PRIS Lab at Beijing University of Posts and Telecommunications [12], runs: BUPTPRISZHS
- SetuServ [6], runs: SS*
- Stanford University, runs: INFLBSTF
- University of Iowa [3], runs: UIowaS*
- University of Oxford and University of Southampton [8], runs: Orc*
- York University [5], runs: york*

Table 4 shows the top runs from each group ordered by their LAM score (lower LAM is better). Table 5 shows the top runs from each group ordered by the AUC score (higher AUC is better).

2.5 Discussion

While the best runs on the LAM and AUC measures look very good with high true positive rates and low false positive rates, the LAM and AUC measures do not appear to be ideally suited for evaluating relevance judgments for their ability to evaluate retrieval runs. Of note, several groups' best LAM score occurred on their runs with low false positive rates while their true positive rates suffered. In the case of the run INFLB2012, it received a lower LAM than 3 other groups even though the run only had an average true positive rate of less than 0.04 and a correspondingly high average false negative rate of 0.964. While a low false positive rate is beneficial, such a high false negative rate may make evaluation of retrieval runs difficult.

On further investigation, we found that this apparent issue with LAM was not with the measure itself but with our smoothing of the FPR and FNR rates. On the advice of Charles Clarke, we modified the smoothing to be proportional to the number of relevant and non-relevant documents:

$$fpr = \frac{|FP| + 0.5(1 - R/N)}{|FP| + |TN| + (1 - R/N)}, \quad (6)$$

$$fnr = \frac{|FN| + 0.5(R/N)}{|FN| + |TP| + R/N}, \quad (7)$$

where R is the number of relevant documents and N is the total number of documents. When we make this change to the smoothing, the non-official results shown in Table 6 seem more reasonable.

While the improved smoothing methodology of Equations 6 and 7 helps LAM better reflect performance, LAM may not tell us which relevance assessing process leads to the best evaluation of a set of retrieval runs. We leave for future work a deeper investigation of LAM's utility for evaluation of crowdsourced relevance judgments.

We mentioned earlier that during adjudication, topic 432 was one of the more difficult topics to adjudicate. It turns out that when we computed the average LAM per topic across all submitted runs using the improved smoothing of rates, topic 432 had an average LAM of 0.34, and the topic with the next greatest LAM was topic 420 with a LAM of 0.25. For the crowdsourcing runs, the easiest topic was topic 416 with a LAM of 0.16. Based on the average LAM and our experience with adjudicating it, topic 432 might not be adequately described by its description and narrative.

3 Image Relevance Assessing Task (IRAT)

3.1 Introduction

The Image Relevance Assessment Task (IRAT) was one of the two tasks of this year's crowdsourcing track. The challenge of this task was for participants to crowdsource high quality relevance judgments for 20k topic-image pairs. In addition to the task given to participants, we were interested in exploring the relationship between judgment quality and the ratio of relevant and non-relevant images for a given topic's assessment pool: Can we observe differences in the quality of the crowdsourced relevance judgments submitted by the participants across topics with different levels of relevance saturation?

The image labeling task was chosen motivated by its roots in crowdsourcing, e.g., the popular ESP game [10], but with the specific aim to investigate crowdsourcing issues in test collection creation that go beyond textual documents. Unlike typical image labeling tasks, where images are labeled with descriptive concepts, e.g., sky, clouds, birds, the task in IRAT was to gather relevance decisions for a set of 90 search topics and 20k images (together with image captions). This task reflects a standard test collection creation scenario enabling the evaluation of image search systems. A property of the task setup was that oftentimes only the combination of an image and its caption together revealed whether it was relevant to a given query or not. This was aimed to encourage the development of hybrid systems, where caption-based retrieval is augmented with human judgments on a sample of the images.

While we expected this task to draw in new participants due to its appeal in crowdsourcing, only four groups took part in the challenge and, of those, only two groups submitted their results to the track. Despite this, the task resulted in a high quality re-usable test collection, annotated with both NIST and crowdsourced relevance labels. In the following, we detail the creation of the test collection and the evaluation results of the submitted runs.

3.2 The IRAT Test Collection

To build the IRAT test corpus, we partnered with ImageCLEF and made use of two image collections that were previously or presently used by the ImageCLEF evaluation forum. These collections were selected based on similarities with the relevance assessing task setup of IRAT. In particular, we made use of subsets of the BELGA corpus from ImageCLEF 2009 [7] and the MIRFLICKR corpus from ImageCLEF 2012 [9]. The complete BELGA test collection consists of around 500,000 images with associated captions, provided by the BELGA news agency, 50 diversity test topics, each with several subtopics, and subtopic level relevance judgments that were obtained by the the ImageCLEF 2009 organizers via crowdsourcing. The MIRFLICKR corpus contains about one million images for which 42 test topics were created by the ImageCLEF 2012 organizers in the context of the Concept Retrieval subtask[1].

From the original 50 diversity topics of the BELGA test set (with 206 subtopics), we selected 70 subtopics (from 29 topics) that could be used in a standalone manner. Furthermore, to allow us to study the effects of relevance saturation in the assessment pools on the quality of crowdsourced relevance judgments, the 70 subtopics were selected taking into account the distribution of the relevance labels in the ImageCLEF 2009 qrels. We defined 7 buckets of relevance saturation, varying between 20% to 80% relevant, in steps of 10%. Then, for each topic, we selected 200 images from the ImageCLEF 2009 qrels using a stratified sampling approach, such that we ended up with 10 topics in each bucket, e.g., in the 20% bucket we have 10 topics, each with around 40 relevant images out of the total 200. Due to errors in the crowdsourced qrels, the selection process was guided by manual inspection and adjudication by the organiser, Gabriella Kazai. One topic (topic 59) had to be removed due to errors in the qrels. The selected images, plus an additional up to 30 images per topic were then judged by one or two NIST assessors (15 of the 69 topics were judged by two assessors). We will refer to the judgment sets as A1 and A2; the topics judged only by a single NIST assessor are added to both A1 and A2. The inter-assessor agreement measured using Cohen's Kappa for the 15 double judged topics is 0.77 ($p < 0.001$). The raw agreement, calculated as the ratio of the number of matching judgments and total judgments is 88.73%. Using the NIST judgments as the ground-truth, the final set of 200 images per topic were then selected into the IRAT test collection, staying as close as possible to the original stratified sample distributions. However, due to disagreements between the NIST and the original crowdsourced relevance labels by ImageCLEF 2009, a couple of topics fell into different buckets. The final number of topics in the different buckets is shown in Table 7 for both the A1 and A2 ground-truth sets.

Rel% bucket:	**20**	**30**	**40**	**50**	**60**	**70**	**80**
#Topics in A1:	9	10	10	10	9	10	11
#Topics in A2:	9	9	9	11	10	10	11

Table 7: Distribution of topics across the relevance saturation buckets for the 69 IRAT topics selected from the ImageCLEF 2009 BELGA test set

```
<topic>
<num> 64.47.4 </num>
<title> ronaldo milan </title>
<description> Relevant images will show
photographs of Ronaldo Nazario when he
played in Milan. Images of Ronaldo in any
other teams are irrelevant. Images of
Ronaldo with other people are relevant if
he is shown in the foreground. Images of
Ronaldo in the background are irrelevant.
</description>
<image> belga30/06104494.jpg </image>
</topic>
```

Figure 1: Topic 20001 from the track's test set

An example topic is shown in Figure 1. As it can be seen, for each BELGA topic an example relevant image was included in the topic description.

Another 70 subtopics were selected and distributed with their full set of IamgeCLEF 2009 qrels to the participants as training set.

In addition to the 69 topics from ImageCLEF 2009, we picked 20 topics out of the 42 topics created in this year's ImageCLEF campaign. Each of these topics contained three example images in the topic description. For each of these topics, 300 images were selected from the pooled submission runs of the Im-

[1] `http://imageclef.org/2012/photo-flickr`

ageCLEF 2012 participants by the ImageCLEF organizers and added into the IRAT test collection that was then distributed to the crowdsourcing track participants. However, out of the 300 images per topic, only 230 were judged by NIST assessors (19 topics by two assessor and one topic by one assessor only). We will use the same convention and refer to the NIST ground-truth sets as A1 and A2. The Cohen's Kappa agreement between the two NIST judges is 0.82 ($p < 0.001$) and the raw agreement ratio is 94.98%. No training set was provided from the ImageCLEF 2012 data set.

3.3 Submissions

Participating groups were required to submit runs, each containing at least one relevance label per sample for all the 19,800 topic-image pairs in the distributed test set (69 topics x 200 images, plus 20 topics x 300 images). All crowdsourced or automatically derived labels had to be submitted. One of the submitted runs had to be designated as the primary run. As in TRAT, participating groups had to submit a binary relevance judgment for every image and an optional probability of relevance value (1.0 means relevant and 0.0 means non-relevant).

Two groups submitted runs: SetuServ submitted 4 runs and UAustin contributed one run. All the runs contained exactly one label per topic-image pair, all of which was labelled as automatically derived. This either means that the teams did not crowdsource any labels or that they omitted those labels from the submission, keeping only the final label per sample. Either way, as a result, we do not have any worker information upon which additional worker reliability analysis could have been performed. Thus, our evaluation in the next section focuses on the accuracy of the submitted labels only.

3.4 Evaluation Results

Table 8 shows the evaluation results for the two primary runs submitted by SetuServ and UAustin. In addition to the metrics discussed in the TRAT section, i.e., LAM and AUC, we also report the number of true positive (TP), true negative (TN), false positive (FP) and false negative (FN) classifications, precision (P), recall (R), accuracy (Acc) and specificity (Spec), where:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN},$$

$$P = \frac{TP}{TP + FP},$$

$$R = \frac{TP}{TP + FN},$$

$$Spec = \frac{TN}{TN + FP}.$$

All evaluation results are calculated over the two judgment sets A1 and A2, both containing all topics that were judged by a single NIST assessor and then A1 including one set of judgments for the double-judged topics and A2 including the other judgment set. The min/max/avg results are the worst/best/average performances calculated by picking topics with min/max accuracy (Acc) over the A1 and A2 judgment sets. Note that for the 20 ImageCLEF 2012 topics, we ignored the images in the submissions that were not judged by NIST.

3.5 Relevance Saturation and Judgment Quality

Figure 2 shows the distribution of Accuracy and LAM scores (using Equations 4 and 5) over the two primary runs submitted by the two participating groups, calculated over the two ground-truth sets A1 and A2. Both Accuracy and LAM show the same trends: both runs perform better when the assessment pools consist largely relevant or non-relevant images. As the number of relevant and non-relevant images gets closer to equal, the performance decreases. One exception is the 40% bucket, which may have contained relatively easy topics. Thus, if the groups obtained the relevance labels from crowd workers, we may then reason that crowds do better on ”spot the odd one out” type tasks.

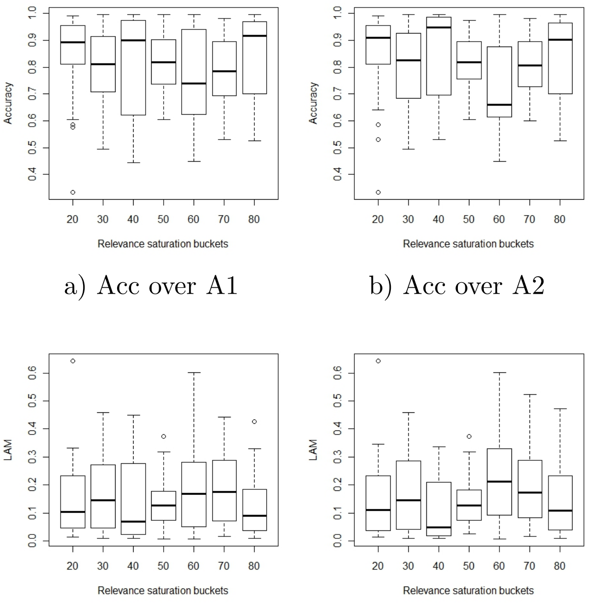
3.6 Conclusions and Future Plans

While the overall performance scores obtained by the participating groups (Accuracy between 0.7-0.9) show already good results, the difference between the runs demonstrates the benefits of different crowdsourcing approaches. In addition, our own investigation of the relationship between judgment quality and relevance saturation levels in the assessment pools suggest that further gains are to be had when relevance saturation can be estimated and the crowdsourcing task tailored accordingly.

Regarding the future of this task, it is unlikely that IRAT will continue again in 2013. However, we will aim to make the dataset available in the future so that crowdsourcing practitioners may use it for research purposes (subject to license clearance). In 2013, we will aim to partner with the TREC Web Track and run a single web page relevance assessing task combining both text and image media.

Table 8: Evaluation results for primary runs (LAM is calculated using Equations 4 and 5, while LAM2 uses Equations 6 and 7

UTAustinM	**TP**	**TN**	**FP**	**FN**	**P**	**R**	**Acc**	**Spec**	**LAM**	**LAM2**	**AUC**
min	63.213	92.483	33.135	17.910	0.641	0.765	0.749	0.712	0.230	0.227	0.528
max	65.472	92.258	30.876	18.135	0.669	0.776	0.759	0.718	0.223	0.218	0.530
avg	64.343	92.371	32.006	18.022	0.655	0.771	0.754	0.715	0.227	0.223	0.529
A1	63.831	92.674	32.517	17.719	0.650	0.771	0.753	0.715	0.226	0.223	0.528
A2	64.854	92.067	31.494	18.326	0.660	0.771	0.755	0.715	0.227	0.222	0.529
SSPostECv2	**TP**	**TN**	**FP**	**FN**	**P**	**R**	**Acc**	**Spec**	**LAM**	**LAM2**	**AUC**
min	72.292	112.011	13.202	9.236	0.817	0.846	0.891	0.877	0.099	0.094	0.865
max	74.708	112.685	10.787	8.562	0.854	0.862	0.906	0.894	0.085	0.080	0.882
avg	73.500	112.348	11.994	8.899	0.836	0.854	0.899	0.885	0.092	0.087	0.873
A1	73.045	112.742	12.449	8.506	0.826	0.855	0.899	0.885	0.092	0.087	0.873
A2	73.955	111.955	11.539	9.292	0.845	0.854	0.899	0.886	0.091	0.086	0.873



a) Acc over A1

b) Acc over A2

c) LAM over A1

d) LAM over A2

Figure 2: Accuracy and LAM (using Equations 4 and 5) score distributions across relevance saturation buckets

4 Acknowledgments

Special thanks to Gaurav Baruah for his on-going and extensive assistance with TRAT technical and non-technical tasks. Le Li was one of the three judges that performed the TRAT adjudication. Special thanks to Bart Thomee at Yahoo! Research who is one of the organizers of ImageCLEF 2012 and who has been instrumental in preparing the Flickr topics and image sets for use in the IRAT. We are grateful for the help of Monica Lestari Paramita, Paul D. Clough, Mark Sanderson and Wuytack Tom for making the BELGA corpus available to us. Thanks to Ellen Voorhees and Ian Soboroff for their help and support with running the track. Thanks to Gordon Cormack for his feedback on the TRAT guidelines. We thank Amazon, CrowdFlower, MobileWorks and Crowd Computing Systems for sponsoring the track and providing credits or discounted prices to track participants.

This work was supported in part by a DARPA Young Faculty Award, in part by a John P. Commons Fellowship, in part by the Natural Sciences and Engineering Research Council of Canada (NSERC), and in part by the University of Waterloo. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsors.

References

[1] M. Bashir, J. Anderton, J. Wu, M. Ekstrand-Abueg, P. B. Golbus, V. Pavlu, and J. A. Aslam. Northeastern university runs at the trec12 crowdsourcing track. In *Online Proceedings of TREC 2012*. NIST, 2013.

[2] G. Cormack and T. Lynam. Trec 2005 spam track overview. In *Proceedings of TREC 2005*. NIST, 2005.

[3] C. Harris and P. Srinivasan. Using hybrid methods for relevance assessment in TREC crowd'12. In *Online Proceedings of TREC 2012*. NIST, 2013.

[4] B. Hedin, S. Tomlinson, J. R. Baron, and D. W. Oard. Overview of the TREC 2009 legal track.

[5] Q. Hu, Z. Xu, Xiangji, and Z. Ye. York university at TREC 2012: Crowdsourcing track. In *Online Proceedings of TREC 2012*. NIST, 2013.

[6] R. Nallapati, S. Peerreddy, and P. Singhal. Skierarchy: Extending the power of crowdsourcing using a hierarchy of domain experts, crowd and machine learning. In *Online Proceedings of TREC 2012*. NIST, 2013.

[7] M. L. Paramita, M. Sanderson, and P. Clough. Diversity in photo retrieval: Overview of the imageclefphoto task 2009. In C. Peters, B. Caputo, J. Gonzalo, G. J. F. Jones, J. Kalpathy-Cramer, H. Müller, and T. Tsikrika, editors, *CLEF (2)*, volume 6242 of *Lecture Notes in Computer Science*, pages 45–59. Springer, 2009.

[8] E. Simpson and S. Reece. Using a Bayesian model to combine LDA features with crowdsourced responses. In *Online Proceedings of TREC 2012*. NIST, 2013.

[9] B. Thomee and A. Popescu. Overview of the imageclef 2012 flickr photo annotation and retrieval task. In P. Forner, J. Karlgren, and C. Womser-Hacker, editors, *CLEF (Online Working Notes/Labs/Workshop)*, 2012.

[10] L. von Ahn and L. Dabbish. Labeling images with a computer game. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, CHI '04, pages 319–326, New York, NY, USA, 2004. ACM.

[11] E. M. Voorhees and D. Harman. Overview of the eigth Text REtrieval Conference (TREC-8). In *Proceedings of the Eighth Text Retrieval Conference (TREC 8)*. NIST, 1999.

[12] C. Zhang, M. Zeng, X. Sang, K. Zhang, and H. Kang. BUPT_PRIS as TREC 2012 crowdsourcing track1:. In *Online Proceedings of TREC 2012*. NIST, 2013.